Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 181-184 recibido 21-03-2018 revisado 18-04-2018 aceptado 03-05-2018

Advanced analytics tool for criminological research of terrorist attacks

Herramienta analítica avanzada para la investigación criminológica de ataques terroristas

Marta Romero Hernández
Pragsis Technologies
Calle de Manuel Tovar 49, Madrid
mromerohdez@gmail.com

Abstract: This paper describes a multilingual text summarization and visual analytics tool that provides a searchable database of over 5,000 different online sources, from news portals to social media links and online documents, related to 10 case studies of terrorism for the period 2013 – 2017.
Keywords: Multilingual text summarization, DANTE, natural language processing, terrorism, visual analytics

Resumen: Este documento describe una herramienta de resúmenes de texto multilingües y visual analytics que proporciona un buscador sobre una base de datos de más de 5.000 fuentes diferentes, desde portales de noticias hasta enlaces de redes sociales y documentos online, relacionados con 10 casos de terrorismo para el período 2013 - 2017.
Palabras clave: Generación de resúmenes multilingües, DANTE, procesamiento del lenguaje natural, terrorismo, visual analytics

1 Introduction and Motivation

The analysis of online resources has become a key issue in the monitoring of terrorist activities. Both a theoretical perspective and a set of methodological tools allow to understand and evaluate terrorist organizations, and to develop anti-terrorism policies and practices to detect and interrupt terrorist attacks.

Total Internet traffic has experienced a dramatic growth in the past two decades. As a result, massive amounts of data are generated every day and shared across different social networks. In particular, content related to terrorism also increases, so the use of powerful filtering tools becomes essential.

This framework has been developed in order to support the criminological analysis of terrorist activities that will be performed in DANTE project. The EC project DANTE "*Detecting and analysing terrorist-related online contents and financing activities*" aims to research and develop technologies for more effective, efficient, automated data mining and analytics. These activities will result in an integrated system to detect, retrieve, collect and analyse huge amounts of heterogeneous and complex multimedia and multi-language terrorist-related contents, from both the Surface and the Deep Web, including Dark nets.

Text summarization represents an essential component of the text analysis services in DANTE. Law Enforcement Agencies (LEAs) spend a lot of time reviewing all the information from internet. By summarizing documents, audio and video transcriptions, officers can easily classify all this content and determine whether they want to explore it in more detail or not.

Given this context, this paper presents a valuable tool that provides visual analytic services to show and trace terrorist activities, terrorist group profiles, etc., in order to reduce the information overload on web intelligence experts due to automated summarization of the relevant content. The proposed application works for English, Italian, Spanish, Portuguese and French and is based on Named Entity Recognition and extractive summarization. The

ISSN 1135-5948. DOI 10.26342/2018-61-27

results of the analysis are presented in a visual way in order to establish complex relationships.

2 Corpus extraction

The domains of interest for the research were the ones already selected by the DANTE Project, namely: i) online financing; ii) online propaganda and iii) online training and information sharing. Comprehensive desk research was carried out along with an in-depth analysis of 10 case studies of terrorist-related activities which have occurred in Europe between 2013 and 2016 (2 cases of foreign fighters; 1 case of failed terrorist attacks; 6 cases of successful terrorist attacks).

From a list of keywords in different languages corresponding to each of these use cases, articles and papers from multiple sources have been retrieved using web scrapping techniques.

The data contains the text for a total of 5024 documents, obtained by public web scraping from the following sources indicated in Table 1:

Source	Language	N. of documents
Reuters	English	716
New York Times	English	119
Perspectives on Terrorism Journal	English	36
La Repubblica	Italian	358
Corriere della Sera	Italian	265
Itistime Universita Cattolica di Milano	Italian	16
El Pais	Spanish	438
El Mundo	Spanish	184
Agencia EFE	Spanish	133
Le Figaro	French	1937
Le Monde	French	617
Le Parisien	French	205

Table 1: Sources from where the data was obtained.

3 Text preprocessing

3.1 Extractive Summarizer

3.1.1 About Text Summarization

There are various kinds of summaries. One distinguishes extractive summaries from abstractive summaries. Extractive summarizers are quite robust since they use existing natural-language phrases that are taken straight from the input. They give a general idea of the text. Abstractive summarizers enable to make more fluent and natural summaries but are more complex to generate and domain-dependent.

There is another classification of automatic summaries that is based on content. An indicative summary is one that informs the reader what a given text or set of texts is about. An informative summary, on the other hand, is one that reproduces the main information of the original and can be used as a replacement of it. Indicative summary contains the main topics of the original text, in contrast to the informative summary that includes full information of the document.

The aim of the summarization task in DANTE is to significantly reduce the text length, without missing the key points of the overall meaning. Since DANTE must deal with arbitrary domain documents (from blogs to propaganda manuals), the indicative-extractive approach, based on term frequency, is selected in this work. This approach is preferred, due to its relative implementation simplicity that is, however, quite robust, unsupervised and fast.

3.1.2 Techniques

Due to the nature of the project, the extractive summarizer is based on term frequency, which has been demonstrated to perform well across several domains. This approach is quite robust since it uses existing natural-language phrases that are taken straight from the input. In addition, it is fast, unsupervised and simple to implement.

The approach presented in this paper consists of three phases:

Phase I: Preprocessing and weighting. The pre-processing step consists in a form of dimensionality reduction by removing noise (e.g. uninformative words or conjugation). This phase involves: splitting the text into segments (phrases, sentences, paragraphs); splitting segments into words (tokenization); word normalization (stemming); stop word filtering and redundancy removal.

After the pre-processing stage, each sentence is scored by the frequency of all of its words, that is, the number of times a word appears in the document. If a word's frequency in a document is high, then it can be assumed that this word has a significant effect on the content of the document. Words occurring in a frequent basis increase the score of their

belonging sentences. The total frequency value of a sentence is calculated by summing up the frequency of every word in the document. Thanks to the pre-processing done previously (elimination of redundant phrases, stop words removal and stemming) we can affirm that more frequent words are indeed significant.

Phase 2: Extraction. After each sentence is scored, they are arranged in descending order of their score value i.e. the sentence whose score value is highest is in top position and the sentence whose score value is lowest is in bottom position. After ranking the sentences based on their total score the summary is produced selecting certain number of top ranked sentences where the required number of sentences is provided by the user.

Phase III: Generation. The algorithm's output was the list of important sentences sorted by score in descending order. The method shows better results if the target number of top sentences is low (2-3 sentences).

3.2 Factual Summarizer

3.2.1 About Named Entity Recognition

Factual summarizer is based on Named Entity Recognition (NER) answers who, where and when of each event in order to generate a global database of events and involved entities.

This summarizer extracts the most significant entities (person, organization, location) from unstructured textual data, to fill event templates and allow browsing document collections according to them.

3.2.2 Techniques

In a preliminary stage a Knowledge Base that contains the known Named Entities is built. This will be used as a dictionary in order to normalize and disambiguate all the different Named Entities recognized by the algorithm.

This database consists of multiple entities classified as LOCATION, PERSON and ORGANIZATION. There is a number of knowledge bases that provide such a background repository for entity classification, predominantly DBpedia, YAGO, and Wikidata. However, there are several reasons to choose Wikidata over other KBs. First, especially when dealing with news articles and social media data streams, it is crucial to have an up-to-date repository of persons and organizations. To the best of our knowledge, Wikidata was the most recent of all, as it provides a weekly data dump. Even though all three KBs (Wikidata, DBpedia, and YAGO3) are based on Wikipedia, Wikidata also contains information about entities and relationships that have not been simply extracted from Wikipedia (YAGO and DBpedia extract data predominantly from infoboxes) but collaboratively added by users.

After the knowledge base is built a process of two phases is performed:

Phase I: Preprocessing. This phase is pretty similar to the extractive summarizer, where the text is splitted in sentences and tokenized. The most important step in this phase is the part-of-speech tagger, or POS-tagger, which processes a sequence of words, and attaches a part of speech tag to each word.

Phase II: Disambiguation rules. The main idea of this point is to "standardize" the named entities using chunk-joining rules such as the location of the entities within the sentence, the positions of uppercase words, the number of repetitions in the text, etc. Then, candidate named entities are searched in the knowledge base where all the named entities are located and used as a dictionary where they are replaced by the "standardized" named entities.

4 Visual representation

The multilingual text summarization and visual analytics tool will allow LEAs to quickly identify multiple sources of data in multiple languages related to an incident and filter the data down to the most relevant information.

The tool has been adapted to provide a separate dashboard for each case study, allowing the user to search through all online content related to a particular case study.

Users may then apply multiple filters using: keywords, news source, location of source, release date and type of source (Figure 1). For analytical purposes the data can then be displayed in graphs to identify peaks in user engagement with a particular news source (Figure 2).

The use of keywords and various filters makes the tool an ideal aid for the methodological approach adopted in DANTE. As indicated in Figure 2, the chronological ordering of articles and the use of keyword searches will make it possible for investigators and researchers similar to quickly construct a crime script of a particular event.

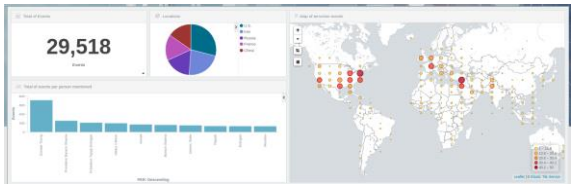

Figure 1: Screenshot of the tool: displaying geographical location of sources on the right and comparative data on the left

Figure 2: Screenshot of the tool: the graph indicates peaks in user engagement with source material and below the title of articles and the summary

5 Implementation

The proposed techniques and methodologies are written in python using the natural language processing toolkit (NLTK).

The last phase of data processing consists of the ingestion in the ELK stack. ELK stands for Elasticsearch, Logstash and Kibana which are technologies for creating visualizations from raw data.

Elasticsearch is an open source, distributed, RESTful search engine, usable by any language that speaks JSON and HTTP.

Kibana is a flexible analytics and visualization platform that lets you set up dashboards for real time insight into your Elasticsearch data.

So, data is ingested with Elastic Search and indexes are created to speed up searches. Then, with Kibana is used to design a dashboard that allows to understand and access information in a simple, easier and fast way.

Acknowledgments

The work presented in this paper was supported by the European Commission under contract H2020-700367 DANTE.

References

Chen, D., J. Bolton, and D.C. Manning. 2016. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2367.

Erkan, G., and D.R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22(1):457–479.

Haghighi, A., and L. Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370.

Lin, C.Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, Vol. 8.

Luhn, P.H. 1958. Automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2): 159-165.

Mani, I., G. Klein, D. House, L. Hirschman, T. Firmin, and B. Sundheim. 2002. Summac: a text summarization evaluation. In *Natural Language Engineering*, 8(1):43–68.

Mihalcea, R., and P. Tarau. 2004. TextRank: Bringing Order into Texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 404-411.

Nallapati, R., B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *CoNLL*.

Nenkova, A., and K. McKeown. 2011. Automatic summarization. In *Foundations and Trends in Information Retrieval*, 5(2-3): 103–233.

Page, L., S. Brin, R. Motwani, and T. Winograd. 1999. The PageRank citation ranking: Bringing order to the web. *Stanford University Technical Report*.

Vanderwende, L., H. Suzuki, C. Brockett, and A. Nenkova. 2007. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. In *Information Processing and Management*, 43(6): 1606-1618.